

(19) World Intellectual Property Organization
International Bureau



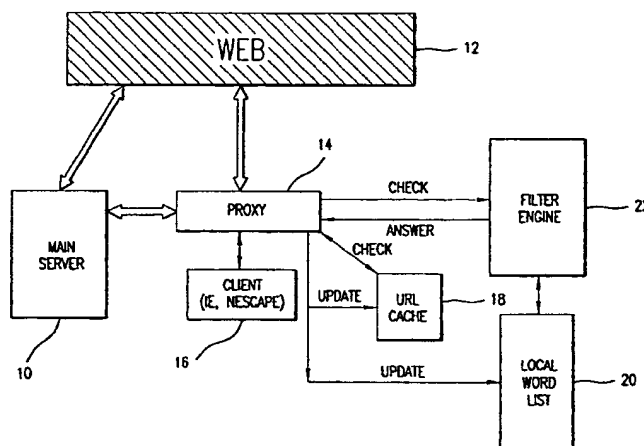
(43) International Publication Date
30 August 2001 (30.08.2001)

PCT

(10) International Publication Number
WO 01/63835 A1

- (51) International Patent Classification⁷: **H04L 9/36**
- (21) International Application Number: PCT/US01/05541
- (22) International Filing Date: 21 February 2001 (21.02.2001)
- (25) Filing Language: English
- (26) Publication Language: English
- (30) Priority Data:
60/183,727 21 February 2000 (21.02.2000) US
60/183,728 21 February 2000 (21.02.2000) US
- (71) Applicant: **CLICKSAFE.COM LLC** [US/US]; 40 East Reading Road, Edison, NJ 08817 (US).
- (72) Inventor: **LIANG, Yufeng**; B3 Sutton Drive, Matawan, NJ 07747 (US).
- (74) Agents: **MORRIS, Francis, E.** et al.; Pennie & Edmonds LLP, 1155 Avenue of the Americas, New York, NY 10036 (US).
- (81) Designated States (*national*): AE, AG, AL, AM, AT, AU, AZ, BA, BB, BG, BR, BY, BZ, CA, CH, CN, CR, CU, CZ, DE, DK, DM, DZ, EE, ES, FI, GB, GD, GE, GH, GM, HR, HU, ID, IL, IN, IS, JP, KE, KG, KP, KR, KZ, LC, LK, LR, LS, LT, LU, LV, MA, MD, MG, MK, MN, MW, MX, MZ, NO, NZ, PL, PT, RO, RU, SD, SE, SG, SI, SK, SL, TJ, TM, TR, TT, TZ, UA, UG, UZ, VN, YU, ZA, ZW.
- (84) Designated States (*regional*): ARIPO patent (GH, GM, KE, LS, MW, MZ, SD, SL, SZ, TZ, UG, ZW), Eurasian patent (AM, AZ, BY, KG, KZ, MD, RU, TJ, TM), European patent (AT, BE, CH, CY, DE, DK, ES, FI, FR, GB, GR, IE, IT, LU, MC, NL, PT, SE, TR), OAPI patent (BF, BJ, CF, CG, CI, CM, GA, GN, GW, ML, MR, NE, SN, TD, TG).
- Published:**
— with international search report
— before the expiration of the time limit for amending the claims and to be republished in the event of receipt of amendments
- For two-letter codes and other abbreviations, refer to the "Guidance Notes on Codes and Abbreviations" appearing at the beginning of each regular issue of the PCT Gazette.

(54) Title: SYSTEM AND METHOD FOR IDENTIFYING AND BLOCKING PORNOGRAPHIC AND OTHER WEB CONTENT ON THE INTERNET



(57) **Abstract:** A system and method are disclosed for identifying and blocking unacceptable web content, including pornographic web content. In a preferred embodiment, the system comprises a proxy server (14) connected between a client (16) and the Internet that checks a requested URL against a block list (18) that may include URLs identified by a web spider. The proxy server requests if the URL is not on the block list, the web content. When the web content is received, the proxy server processes its text content by using filterengine (22) and compares the processing results using a threshold (508). If necessary, the proxy server then processes the image content of the retrieved web content to determine if it comprises skin tones and textures by using texture filter (712) and tone filter (710). Based on these processing results, the proxy server may either block the retrieved web content or permit user access to it. Also disclosed is a system and method for inserting advertisements into retrieved web content.

SYSTEM AND METHOD FOR IDENTIFYING AND BLOCKING PORNOGRAPHIC AND OTHER WEB CONTENT ON THE INTERNET

This application claims priority to United States Provisional Application No. 60/183,727 and United States Provisional Application No. 60/183,728, each of which is hereby
5 incorporated by reference.

Background of the Invention

Tools for identifying and blocking pornographic websites on the Internet are known in
10 the art. Typically, these tools comprise a "block" list comprising URLs of known pornographic sites. When an unauthorized user attempts to retrieve web content from a site on the block list, the user's browser blocks the request.

It is difficult, however, to keep the block list current because objectionable web sites are constantly being added to the Internet. Moreover, these prior art tools fail to block sites
15 that are not on the block list.

Summary of the Invention

A system and method are disclosed for identifying and blocking unacceptable web content, including pornographic web content. In a preferred embodiment, the system
20 comprises a proxy server connected between a client and the Internet that processes requests for web content. The proxy server checks the requested URL against a block list that may include URLs identified by a web spider. If the URL is not on the block list, the proxy server requests the web content.

When the web content is received, the proxy server processes its text content and
25 compares the processing results using a threshold. If necessary, the proxy server then processes the image content of the retrieved web content to determine if it comprises skin tones and textures. Based on these processing results, the proxy server may either block the retrieved web content or permit user access to it.

Also disclosed is a system and method for inserting advertisements into retrieved web
30 content. In a preferred embodiment, the system inserts html content that may comprise a hyperlink into the top portion of the retrieved web content.

Brief Description of the Drawings

The above summary of the invention will be better understood when taken in conjunction with the following detailed description and accompanying drawings, in which:

Fig. 1 is a block diagram of a first preferred embodiment of the present system;

5 Fig. 2 is a block diagram of a second preferred embodiment of the present system;

Fig. 3 is a flow diagram depicting a preferred process implemented by the embodiments shown in Figs. 1 and 2;

Fig. 4A is a flow diagram depicting a preferred embodiment of a text analysis algorithm employed by the present system;

10 Fig. 4B is a preferred embodiment of a lexicon of words and values assigned to them employed by the present system;

Fig. 5 is a block diagram of a preferred text analysis engine of the present system;

Fig. 6 is a flow diagram depicting a preferred embodiment of an algorithm for determining the h values used by the text analysis engine of Fig. 5;

15 Fig. 7 is a block diagram of a preferred image analysis engine of the present system;

Fig. 8A is a flow diagram depicting a preferred filtering algorithm for use in the present system;

Fig. 8B depicts an image area to be filtered using the filtering algorithm depicted in Fig. 8A;

20 Fig. 9 is a flow chart depicting a preferred algorithm employed by a web spider to create a list of unacceptable web sites; and

Fig. 10 is a flow chart depicting a preferred algorithm for inserting advertisements into retrieved web content.

25 Detailed Description of the Preferred Embodiments

Fig. 1 is a block diagram of a first preferred embodiment of the present system. As shown in Fig. 1, the system preferably comprises a proxy server 14 that is designed to receive URL requests for web content from a client 16. Typically, client 16 will be one of many clients connected to a network (not shown). Each request for web content by a client 16 that
30 is transmitted over the network is forwarded to proxy server 14 for processing.

Proxy server 14 determines whether the request is permissible (as described in more detail below) and, if it is, forwards the request to an appropriate web site (not shown) via world-wide-web 12. When a web page or other content is received from the web site, proxy server 14 determines whether the content is acceptable, and, if it is, forwards the web page to client 16.

In a preferred embodiment, a URL is deemed acceptable if it does not identify a pornographic web site. Similarly, a web page or other web content is acceptable if it does not comprise pornographic content.

As further shown in Fig. 1, the system also preferably comprises a URL cache 18 that stores a list of impermissible URLs. In addition, the system preferably comprises a local word list 20 and a filter engine 22 which are used by proxy server 14 to identify pornographic material, as described in more detail below.

In a preferred embodiment, URL cache 18 may be populated in several ways. First, cache 18 may be populated with a list of known pornographic websites. Second, an authorized user may specify specific URLs that are unacceptable. Third, an authorized user may specify specific URLs that are acceptable (i.e., that should not be blocked, even though the remaining components of the system, described below, would identify the content as pornographic). Fourth, URL cache 18 may be populated by a web spider. A preferred embodiment of a particular web spider for use with the present system is described in more detail below.

In a preferred embodiment, when a site is designated acceptable even though it comprises pornographic material, access to that site is limited to authorized individuals, such as, for example, the individual that designated the site acceptable. In this way, for example, an adult may designate certain sites acceptable and nevertheless block access to such sites by a child.

Also shown in Fig. 1 is a main server 10. Main server 10 serves several functions including maintaining an updated list of unacceptable URLs, as described in more detail below. Typically, main server 10 is not co-located with proxy server 14 or client 16. Rather, it is typically located in a remote location from where it may provide updated unacceptable URL lists and other services to a plurality of proxy servers 14 and clients 16.

Fig. 2 is an alternative preferred embodiment of the present system. As shown in Fig. 2, in this alternative embodiment, a client 16 may be connected directly to the Internet. In that event, URL cache 18, local word list 20, filter engine 22, as well as software 24 for using these modules is preferably resident in client 16.

5 Fig. 3 is a flow diagram depicting a preferred process implemented by the embodiments shown in Figs. 1 and 2. For purposes of ease of description, the following description will refer primarily to the architecture disclosed in Fig. 1. It will be understood, however, that the same steps may be performed by corresponding components shown in Fig. 2. In addition, it should be noted that although the steps in Fig. 3 are demonstrated as
10 sequential, the text and image analysis engines described below may instead be designed to operate in parallel. In particular, parallel operation may be desirable when large processing resources are available, while the serial approach described below may be preferable when there is a desire to conserve processing resources.

Turning to Fig. 3, in step 302, a user enters a URL onto the command line of his or
15 her browser. In step 304, server 14 compares the URL to the list of unacceptable URLs stored in URL cache 18. If the URL is on the list, then server 14 blocks the user's request, and does not obtain the requested web page specified by the URL.

Otherwise, if the URL is acceptable, server 14 transmits a URL request via web 12 to retrieve the requested web page (step 306). When the web page is returned, server 14
20 conducts a text analysis of the text content of the web page (step 308). A preferred embodiment of this text analysis is described in connection with Figs. 4-6.

As shown in Fig. 4A, in step 402, server 14 first analyzes the text content of the retrieved web page and identifies every word or combination of words that it contains. It should be noted that this text search preferably includes not only text that is intended to be
25 displayed to the user, but also html meta-text such as hyperlinks. It should also be noted that the identified words may include a substring within a longer word in the text.

In step 404, server 14 compares each word and combination of words to a lexicon of words stored in local word list 20. A preferred embodiment of lexicon 20 is shown in Fig. 4B.

30 It should be noted that each of the words in the lexicon shown in Fig. 4B has two values following it, and that those words associated with the preferred embodiment being

discussed presently are those that have a "0" as their second value. These words are associated with pornography and are utilized by the system to identify pornographic material, as described below. Words having a value other than "0" as their second value are preferably associated with other concepts or categories of material, as described in more detail below.

5 As further shown in Fig. 4B, each word or combination of words in local word list 20 is also assigned a first value. In the preferred embodiment shown in Fig. 4B, this first value is between 0.25 and 8. If a word or combination of words found in the web content is in the lexicon, server 14 retrieves this assigned value for the word or combination of words.

10 In step 406, server 14 uses the retrieved values as inputs to a text analysis engine for determining a score that is indicative of the likelihood that the retrieved web content is pornographic. In a preferred embodiment, the text analysis engine employs artificial intelligence to determine the likelihood that the retrieved web content is pornographic. A block diagram of a preferred text analysis engine is described in connection with Fig. 5.

15 As shown in Fig. 5, text analysis engine 502 preferably comprises a plurality of inputs x_1, x_2, \dots, x_n which are provided to multipliers 504. Each x_i represents the value retrieved from local word list 20 for the i^{th} word or combination of words found in the text of the retrieved web content. It should be noted that if a word in the lexicon appears n times in the text, the system preferably multiplies the retrieved value assigned to the word by n and supplies this product as input x_i to text analysis engine 502.

20 Each multiplier 504 multiplies one input x_i by a predetermined factor h_i . A preferred method for determining factors h_1, h_2, \dots, h_n is described below.

25 The outputs of multipliers 504 are then added in adder 506. The output of adder 506 is then provided to a thresholder 508 that implements a sigmoid function. The output of thresholder 508 therefore may be: 1) less than a lower threshold; 2) between a lower threshold and an upper threshold; or 3) above the upper threshold. In a preferred embodiment, the lower threshold may be approximately 0.25 and the upper threshold may be approximately 0.5.

30 Returning to step 308 of Fig. 3, if the output of thresholder 508 is below the lower threshold, then server 14 concludes that the retrieved web content is not pornographic, and server 14 forwards the retrieved web content to client 16 (step 310). If the output of thresholder 508 is above the upper threshold, then server 14 concludes that the retrieved web

content is pornographic, and server 14 "blocks" the content by not sending it to client 16 (step 312).

If, however, the output of thresholder 508 is above the lower threshold but below the upper threshold, then the system proceeds to step 314, where it analyzes the image content of the retrieved web content to determine whether the retrieved web content is pornographic.

Before turning to step 314, however, a preferred embodiment for determining the h values used by the text analysis engine is first described in connection with Fig. 6. The steps in this preferred embodiment may, for example, be performed by main server 10.

As shown in Fig. 6, in step 602 a plurality of web sites are shown to a plurality of people. With respect to each web site, each person states whether they consider the site's content to be pornographic or not. In step 604, the text content of each web page categorized by the plurality of people is analyzed to identify every word and combination of words that it contains. In step 606, each word and combination of words is compared to a lexicon of words, typically the same as the lexicon stored in local word list 20. If a word or combination of words found in the web content is in the lexicon, the assigned value for the word or combination of words is retrieved.

In step 608, the system generates an equation for each person's opinion as to each web site. Specifically, the system generates the following set of equations:

$$(x_1^{(1)} * h_1) + (x_2^{(1)} + h_2) + \dots (x_n^{(1)} * h_n) = y_1$$

$$(x_1^{(2)} * h_1) + (x_2^{(2)} + h_2) + \dots (x_n^{(2)} * h_n) = y_2$$

$$(x_1^{(A)} * h_1) + (x_2^{(A)} + h_2) + \dots (x_n^{(A)} * h_n) = y_A$$

OR:

$$[X] * [H] = [Y]$$

where:

x_i is the value retrieved from the database for the i^{th} word or combination of words found in the text of the web site that is also in the lexicon,

5 h_i is the multiplier to be calculated for the i^{th} word or combination of words found in the text of the web site that is also in the lexicon, and

y_j is either 0 or 1 depending on whether the j^{th} person stated that he or she found the web site to be pornographic or not (0 = not pornographic).

In step 610, the system solves this matrix of equations as:

10

$$[H] = [X]^{-1} [Y]$$

It should be noted that when $[X]$ does not have an inverse, a least square algorithm may instead be used as an approximation for the value of $[X]^{-1}$. It should also be noted that if the
15 x values are chosen wisely, then one may expect the h values to fall between 0.9 and 1.1.

Returning to Fig. 3, recall that when the text analysis fails to conclusively demonstrate whether the retrieved web content is or is not pornographic, the system proceeds to step 314 where an image analysis of the retrieved web content is performed. A preferred embodiment for performing this image analysis is described in connection with Fig. 7.

20

Fig. 7 is a block diagram of a preferred image analysis engine of the present system. As shown in Fig. 7, an image analysis engine 702 preferably comprises an adder 704 that receives the luminescence values for the red, green, and blue components of each pixel in the image and adds them to determine brightness ($L=R+G+B$). A first divider 706 divides this sum by the pixel's red value to determine the normalized red value r , where $r = R/(R+G+B)$.
25 Similarly, a second divider 708 divides the brightness by the pixel's blue value to determine the normalized blue value b , where $b = B/(R+G+B)$. Together, these two values, r and b , define the image tone for each pixel.

30

Values r and b are supplied to a tone filter 710. Interestingly, it has been found that although images of human skin appear markedly different to viewers (e.g., white, black, yellow, brown, etc.), this difference is a function of the image brightness rather than the tone. In fact, it has been found that the distribution of pixels representing skin in an image is

relatively constant and follows a Gaussian distribution. Therefore, if the normalized red and blue values of all the pixels in an image are plotted on a graph of r vs. b , approximately 95% of pixels in the image that represent skin will fall within three standard deviations of the intersection of the mean values of r and b for pixels representing skin. Tone filter 710

5 identifies pixels having r and b values within three standard deviations of the mean values of r and b and thus identifies portions of the image that are likely to include skin.

Interestingly, it has been found that areas in an image representing skin typically have relatively low granularity. As a consequence, such areas of the image have little energy in the high spatial frequency. Areas of the image that include skin can therefore be distinguished by

10 a high-pass spatial filter. A preferred embodiment for a texture filter 712 incorporating such a high-pass spatial filter is described in connection with Figs. 8A-B.

Texture filter 712 preferably employs multi-resolution median ring filtering to capture multi-resolution textural structure in the image being considered. A median filter may essentially be considered as a band-pass filter. Median filters are non-linear and, in most

15 cases, are more robust against spiky image noise. Such filters capture edge pixels in multiple resolutions using a recursive algorithm, depicted in Fig. 8A.

As shown in Fig. 8A, in step 802, the filter is set to a first ring radius r . In a preferred embodiment, r may be initially set to 13. In step 804, the image is filtered by replacing each pixel x_k in the image with the median of the values of eight pixels lying on a circle at radius r from pixel x_k , as shown in Fig. 8B for the example of $r=3$. Thus, each pixel x_k is replaced by:

20 $\text{median}(x_0, x_1, x_2, \dots, x_7)$. This process is equivalent to conducting a non-linear band-pass filtering of the image.

In step 806, it is determined whether $r=1$. If it is, then the process finishes at step 808. Otherwise, r is set to $r-1$ (step 810), and the process loops back to step 804 to again filter the

25 image. Thus, filtering is recursively conducted until r is equal to 1.

The resulting image is a smoothed version of the original image at various resolutions. Texture filter 712 then abstracts this resulting image from the original image to obtain the texture image.

Once the texture image is obtained, a local 5 X 5 average "I" of the image is obtained

30 for each pixel (i,j) and that average is compared to a threshold. If $I(i,j) > \text{threshold}$, then (i,j)

is considered to be a textural pixel, and thus does not represent a skin area. Otherwise, if $I(i,j) < \text{threshold}$, then (i,j) is considered not a textural pixel.

The outputs of tone filter 710 and texture filter 712 are ANDed together by logical AND 714. If tone filter 710 identifies a pixel as having a skin tone and texture filter 712
5 identifies a pixel as being a not textural pixel, then the output of logical AND 714 indicates that the pixel represents a skin area.

As noted above, in a preferred embodiment, URL cache 18 may be populated by a web spider 26. Web spider 26 may preferably be co-located with main server 10, and may periodically download to server 14 an updated list 28 of URLs of pornographic web sites that
10 it has compiled. Web spider 26 is preferably provided with a copy of the lexicon described above as well the text analysis engine and image analysis engine described above so as to permit it to recognize pornographic material. A preferred embodiment of a particular web spider for use with the present system is now described in connection with Fig. 9.

As shown in Fig. 9, in step 902, web spider 26 is provided with a first URL of a web
15 site known to contain pornographic material. In a preferred embodiment, the web site is one that comprises a plurality of links to both additional pages at the pornographic website, as well as other pornographic websites.

In step 904, web spider 26 retrieves the web page associated with the first URL. In step 906, web spider 26 determines whether the retrieved web content contains pornographic
20 material. If it does, then in step 908, web spider 26 adds the URL to list 28.

In step 910, web spider 26 then retrieves another web page having a link in the first URL that it received. The process then returns to step 906, where web spider 26 again determines whether the retrieved web page comprises pornographic material and, if it does, to step 908, where the URL of the pornographic page is added to list 28.

25 This loop preferably continues until web spider 26 exhausts all web pages that link, directly or indirectly, to the first URL that it was provided. At that point, an additional "seed" URL may be provided to web spider 26, and the process may continue.

In a preferred embodiment, web spider 26 employs a width-first algorithm to explore all linked web pages. Thus, for example, web spider 26 examines the web pages linked by
30 direct links to the original URL before proceeding to drill down and examine additional pages linked to those pages that link to the original URL.

In a preferred embodiment, if any page in a website is discovered as comprising pornographic material, all pages "below" that page in the sitemap for the web site may be blocked. Pages above the pornographic page may preferably remain unblocked.

5 Alternatively, an entire website may be designated unacceptable if any of its web pages are unacceptable.

In a further preferred embodiment, a user may program the system to filter out additional subject matter that is not, strictly speaking, pornographic. For example, if desired, the system may identify material relating to the concepts "bikini" or "lingerie". In the exemplary lexicon shown in Fig. 4B, for example, the words "lingerie," "bra," etc. are
10 included in the lexicon and assigned a second value equal to "1" to identify them as belonging to the lingerie category. The system will then search for these terms during the text analysis and, either on the basis of text alone, or in combination with the image analysis, will identify and block web content directed to these subjects.

In addition, a user may program the system to filter out subject matter relating to other
15 areas such as hate, cults, or violence by adding terms relating to these concepts to the lexicon. The system will then search for these terms during the text analysis and block web content directed to these subjects. In the exemplary lexicon shown in Fig. 4B, for example, words associated with hate groups may be added to the lexicon and assigned a second value equal to 2, words associated with cults may be added to the lexicon and assigned a second value equal
20 to 3, and words associated with violence may be added to the lexicon and assigned a second value equal to 4. In addition, other words that do not necessarily correspond to a defined category (e.g., marijuana), may be added to the lexicon and assigned a second value equal, e.g., to 5, if they are deemed likely to occur in objectionable material.

In another aspect, the present system may also comprise the capability to insert
25 advertisements into web pages displayed to a user. This preferred embodiment is described in connection with Fig. 10. As shown in Fig. 10, in step 1002, server 14 receives a web page from web 12. In step 1004, server 14 determines whether the content of the web page is acceptable, as described in detail above.

In step 1006, server 14 retrieves from memory an advertisement for insertion into the
30 web page. In a preferred embodiment, this advertisement may include an html link to be inserted near the top of the retrieved html web page.

In step 1008, server 14 inserts the advertisement into the retrieved web content. Thus, for example, after the ad is inserted, the retrieved web content may take the following form:

```
5  <html>
    <head>                                </head>
    <body>
    <a href = "http://www._____.com"> Buy Golf Equipment! </a>
    </body>
    </html>
```

10

In a preferred embodiment, server 14 inserts the advertisement into the top portion of the retrieved web page, even if the retrieved web page comprises several frames. This may be accomplished, for example, with a short piece of Javascript. For example:

```
15  <script.Javascript>
    if (self = top | self = top.frame[0])
    insert (advertisement)
```

20 While the invention has been described in conjunction with specific embodiments, it is evident that numerous alternatives, modifications, and variations will be apparent to those skilled in the art in light of the foregoing description.

Claims

1. A system for identifying possibly pornographic web sites comprising:
a feature extraction module, the feature extraction module comprising:
 - 5 a first module for extracting the URL of the website from a request for web content;
 - a second module for extracting text from text portions of the web page;
 - a third module for extracting image portions from the web page that likely correspond to the skin of an individual; and
 - 10 a fusion module for evaluating the output from the feature extraction module and determining whether the web page comprises possibly pornographic content.
2. The system of claim 1, further comprising a URL cache.
- 15 3. The system of claim 2, wherein the URL cache comprises a list of unacceptable URLs.
4. The system of claim 2, wherein the URL cache comprises a list of acceptable URLs.
5. The system of claim 4, wherein the acceptable URLs are accessible only by authorized
20 individuals.
6. The system of claim 2, wherein the URL cache is populated by a web spider.
7. The system of claim 1, further comprising a list of words found in pornographic material.
25
8. The system of claim 7, wherein each word in the list is assigned a value.
9. The system of claim 8, further comprising a text analysis engine.
- 30 10. The system of claim 9, wherein the text analysis engine multiplies the assigned value for every word on the list that is also in the text portion of a web page by an associated value,

sums together the products, and supplies the sum to a thresholder implementing a sigmoid function.

11. The system of claim further comprising an image analysis engine.

5

12. The system of claim 11, further comprising a tone filter.

13. The system of claim 11, further comprising a texture filter.

10 14. A method for inserting an advertisement into retrieved web content, comprising:
retrieving web content;
retrieving an advertisement;
inserting the advertisement into the web content in a computer that is either the client
computer that requested the web content or a server connected to the same LAN or WAN as
15 the computer that requested the web content.

15. The method of claim 14, wherein the advertisement comprises html content.

16. The method of claim 14, further comprising the step of checking the web content to
20 determine if it is pornographic before permitting the web content to be displayed to a user.

1/15

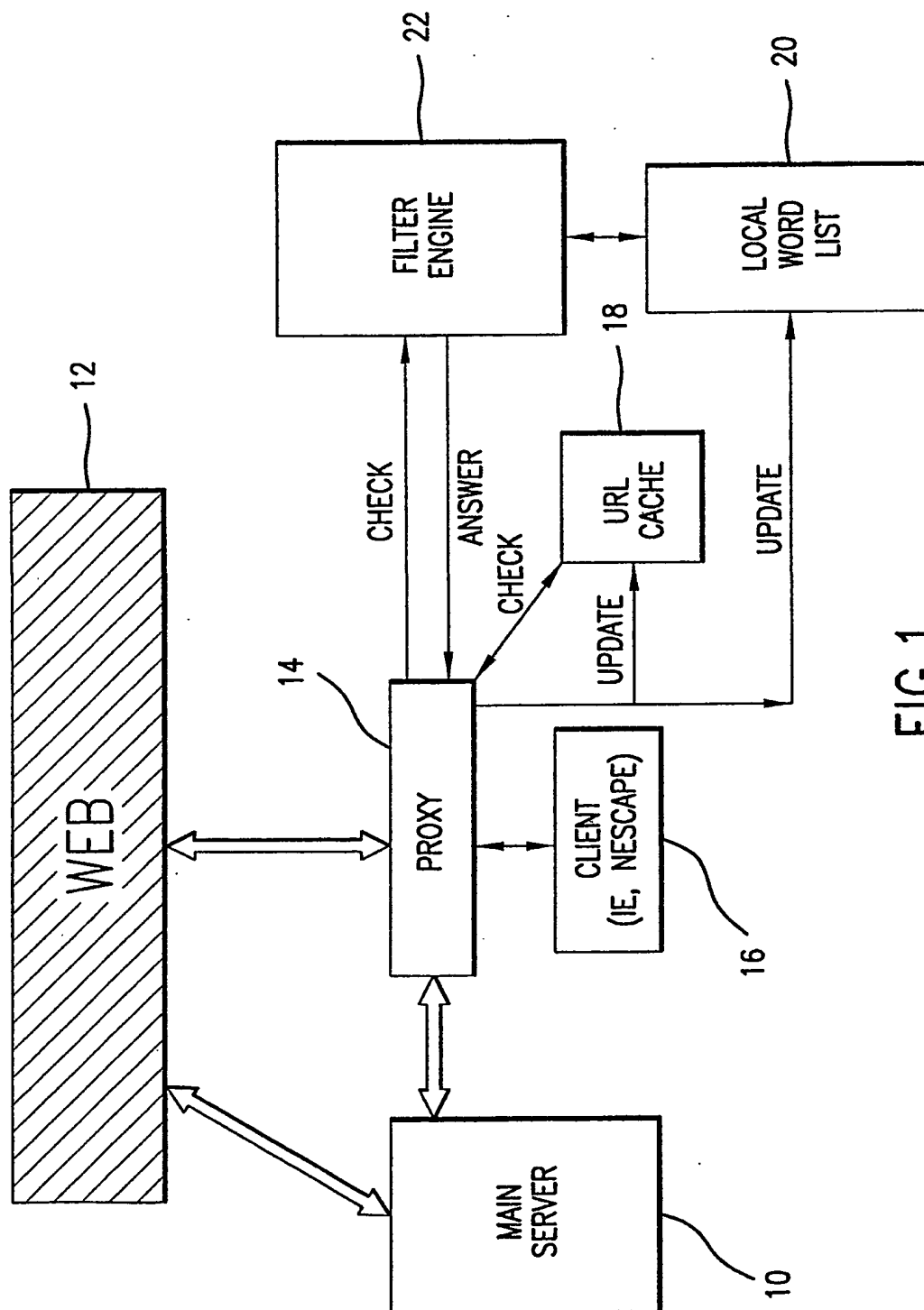


FIG.1

2/15

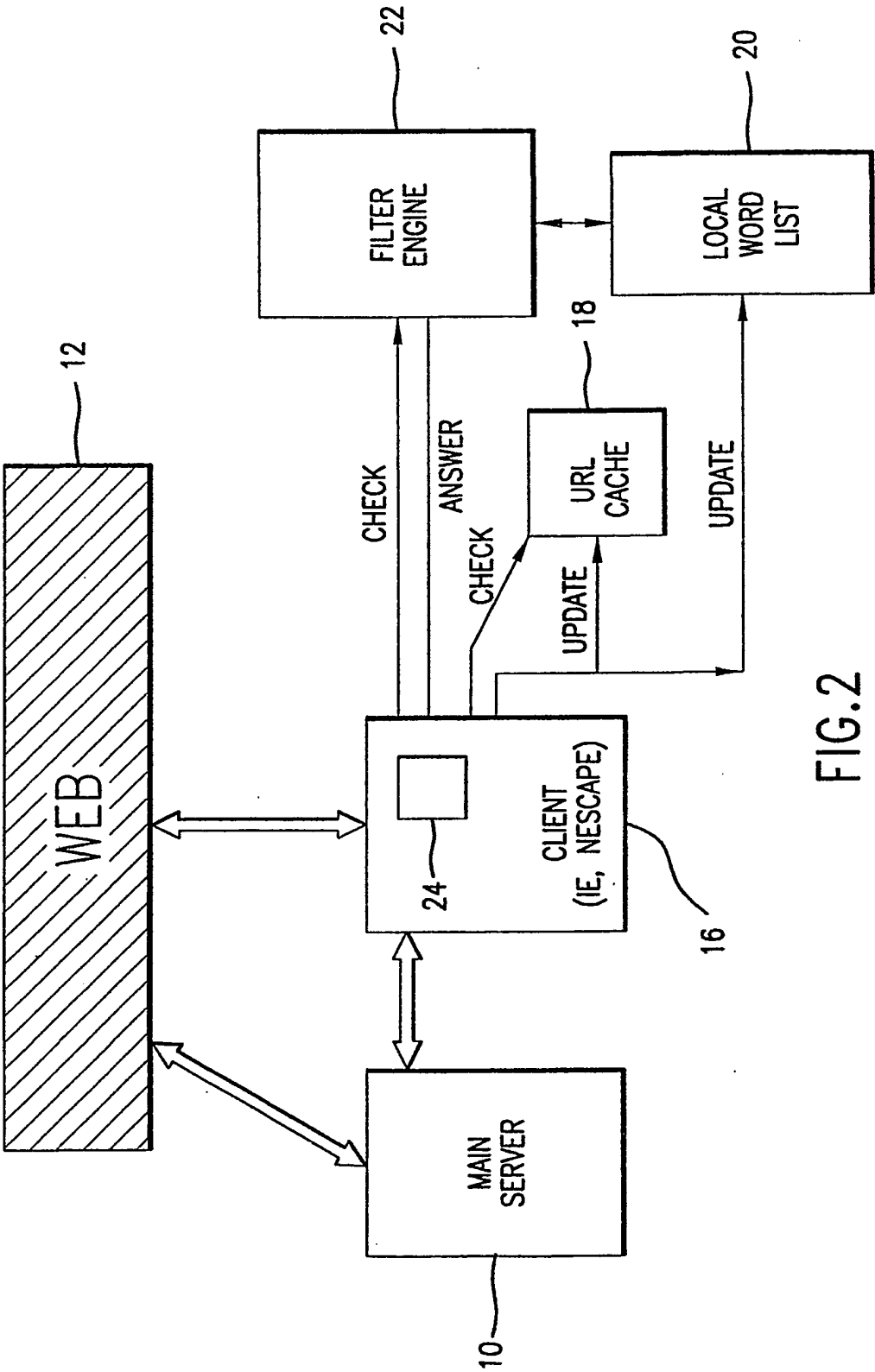


FIG.2

3/15

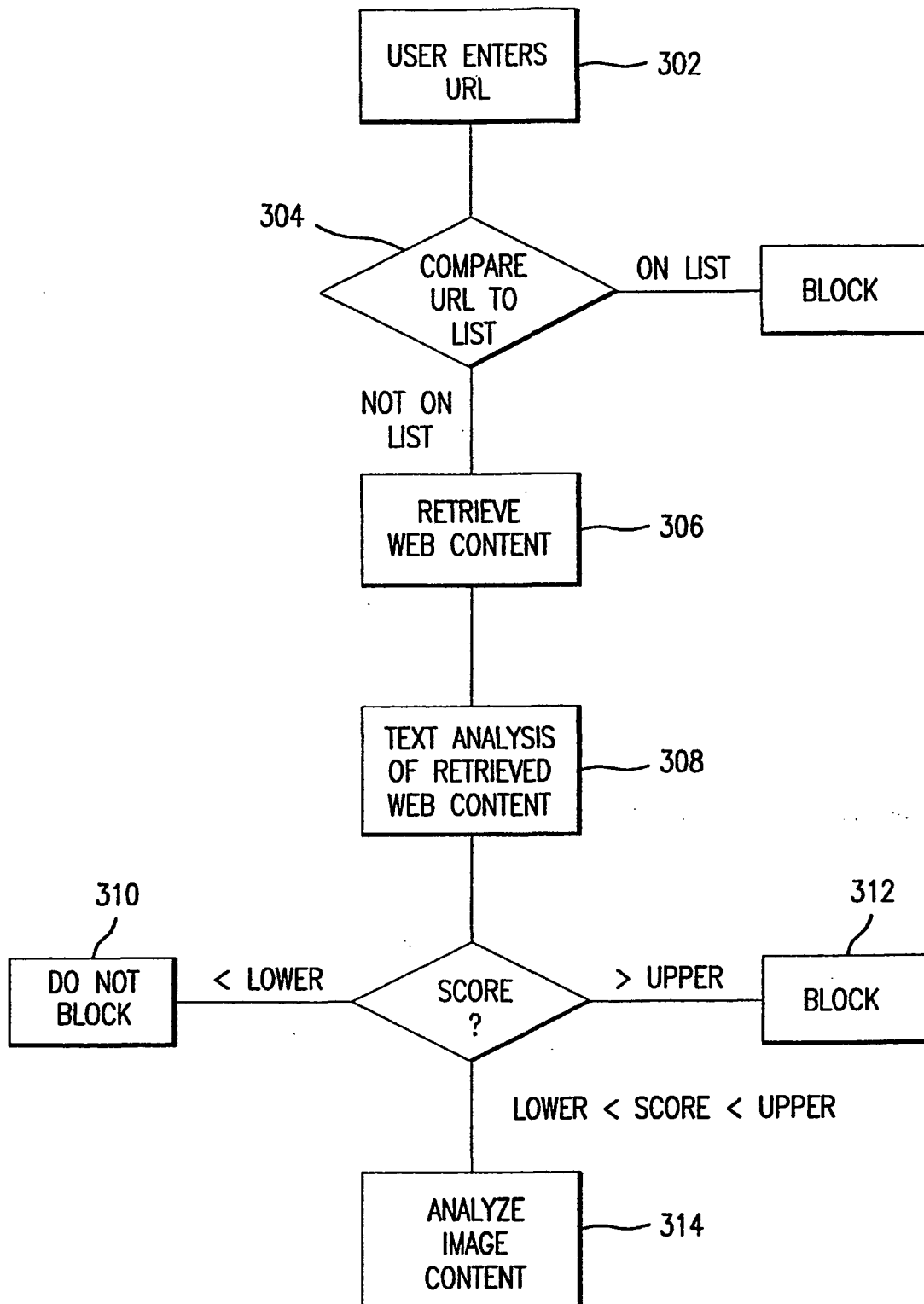


FIG.3

SUBSTITUTE SHEET (RULE 26)

4/15

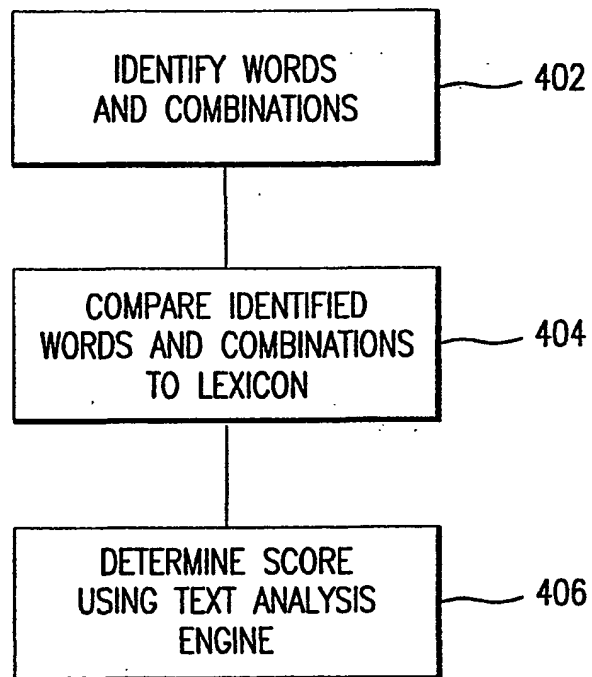


FIG. 4A

5/15

"areola", 0.12, 0
"clitoris", 2.00, 0
"condom", 0.12, 0
"diaphragm", 0.12, 0
"dominance", 0.12, 0
"E-zines", 8.00, 0
"foreskin", 0.12, 0
"genitalia", 0.25, 0
"hymen", 2.00, 0
"kinky", 4.00, 0
"labia", 2.00, 0
"nasty", 0.50, 0
"seductive", 0.50, 0
"submission", 0.12, 0
"swinging", 0.50, 0
"urine", 0.12, 0
"urination", 0.50, 0
"adultcheck", 8.00, 0
"adultsights", 8.00, 0
"anal", 8.00, 0
"analingus", 8.00, 0
"ass", 2.00, 0
"asshole", 8.00, 0
"bestiality", 8.00, 0
"bestial", 8.00, 0
"bestiality", 8.00, 0
"bisexual", 1.00, 0
"blowjob", 8.00, 0
"blowjobs", 8.00, 0
"bomb", 0.12, 0
"bondage", 8.00, 0
"boob", 2.00, 0
"buttfucking", 8.00, 0
"cannibalism", 8.00, 0
"clit", 2.00, 0
"cock", 8.00, 0
"cocks", 8.00, 0
"coitus", 8.00, 0
"copulate", 4.00, 0
"copulation", 4.00, 0
"cum", 8.00, 0
"cumshot", 8.00, 0
"cumshots", 8.00, 0
"cunnilingus", 8.00, 0

FIG.4B-1

SUBSTITUTE SHEET (RULE 26)

6/15

"cunt", 8.00, 0
"cunts", 8.00, 0
"decadence", 2.00, 0
"dicks", 8.00, 0
"dildo", 8.00, 0
"dildos", 8.00, 0
"doobie", 8.00, 0
"drugs", 0.12, 0
"ejaculate", 2.00, 0
"ejaculation", 2.00, 0
"erection", 4.00, 0
"erotic", 8.00, 0
"erotica", 8.00, 0
"exhibitionism", 8.00, 0
"exhibitionist", 8.00, 0
"exhibitionists", 8.00, 0
"felching", 8.00, 0
"fellatio", 8.00, 0
"fetish", 2.00, 0
"fetishes", 2.00, 0
"fistfuck", 8.00, 0
"fisting", 8.00, 0
"flesh", 1.00, 0
"frottage", 8.00, 0
"fuck", 8.00, 0
"fucked", 8.00, 0
"fuckers", 8.00, 0
"fucking", 8.00, 0
"gangbang", 8.00, 0
"gerbiling", 8.00, 0
"groupsex", 8.00, 0
"hard-on", 8.00, 0
"hardcore", 8.00, 0
"hardon", 8.00, 0
"heterosexual", 1.00, 0
"homosexual", 1.00, 0
"horniest", 4.00, 0
"horny", 4.00, 0
"incest", 2.00, 0
"intercourse", 8.00, 0
"jism", 8.00, 0
"kinky", 8.00, 0
"lesbian", 1.00, 0
"lezbos", 2.00, 0
"lusting", 2.00, 0

FIG.4B-2

SUBSTITUTE SHEET (RULE 26)

7/15

"masochism", 8.00, 0
"masturbate", 2.00, 0
"masturbation", 2.00, 0
"nude", 4.00, 0
"nudes", 4.00, 0
"nudity", 4.00, 0
"nympho", 8.00, 0
"nymphomania", 8.00, 0
"nymphomaniac", 8.00, 0
"obsex", 8.00, 0
"orgasm", 4.00, 0
"orgy", 8.00, 0
"penis", 2.00, 0
"perverse", 1.00, 0
"perversion", 1.00, 0
"perverted", 1.00, 0
"porn", 0.50, 0
"porno", 0.50, 0
"pornography", 0.50, 0
"prick", 1.00, 0
"prostitution", 0.50, 0
"pussies", 8.00, 0
"pussy", 8.00, 0
"rape", 0.50, 0
"rimming", 0.12, 0
"sadism", 8.00, 0
"sadoomasochism", 8.00, 0
"s&m", 8.00, 0
"s/m", 8.00, 0
"screwing", 8.00, 0
"sexy", 0.25, 0
"sexual", 0.25, 0
"shemales", 4.00, 0
"slut", 2.00, 0
"sluts", 2.00, 0
"smut", 4.00, 0
"snatch", 0.12, 0
"snatches", 0.12, 0
"sodomy", 2.00, 0
"spank", 4.00, 0
"spunk", 8.00, 0
"suck", 2.00, 0
"threesome", 0.25, 0
"tit", 4.00, 0
"tits", 4.00, 0
"transexuality", 4.00, 0

FIG.4B-3

SUBSTITUTE SHEET (RULE 26)

8/15

"transvestite", 4.00, 0
"twat", 8.00, 0
"vibrator", 8.00, 0
"voyeur", 8.00, 0
"voyeurism", 8.00, 0
"vulva", 8.00, 0
"whore", 8.00, 0
"xxx", 8.00, 0
"zoophile", 8.00, 0
"zoophilia", 8.00, 0
"asb", 8.00, 0
"asw", 8.00, 0
"ass", 4.00, 0
"assd", 8.00, 0
"apbe", 8.00, 0
"b&d", 8.00, 0
"bdsm", 8.00, 0
"d&s", 8.00, 0
"motas", 8.00, 0
"motos", 8.00, 0
"motss", 8.00, 0
"sensual", 0.50, 0
"sensuality", 0.50, 0
"lingerie", 4.00, 1
"panty", 4.00, 1
"bra", 1.00, 1
"bras", 1.00, 1
"marijuana", 0.50, 5
"underware", 2.00, 1
"luscious", 2.00, 0
"intimacy", 0.75, 0
"intimate", 0.75, 0
"dominatrix", 2.00, 0
"dominant", 0.50, 0
"dominance", 0.50, 0
"submission", 0.50, 0
"submissive", 0.50, 0

FIG.4B-4

9/15

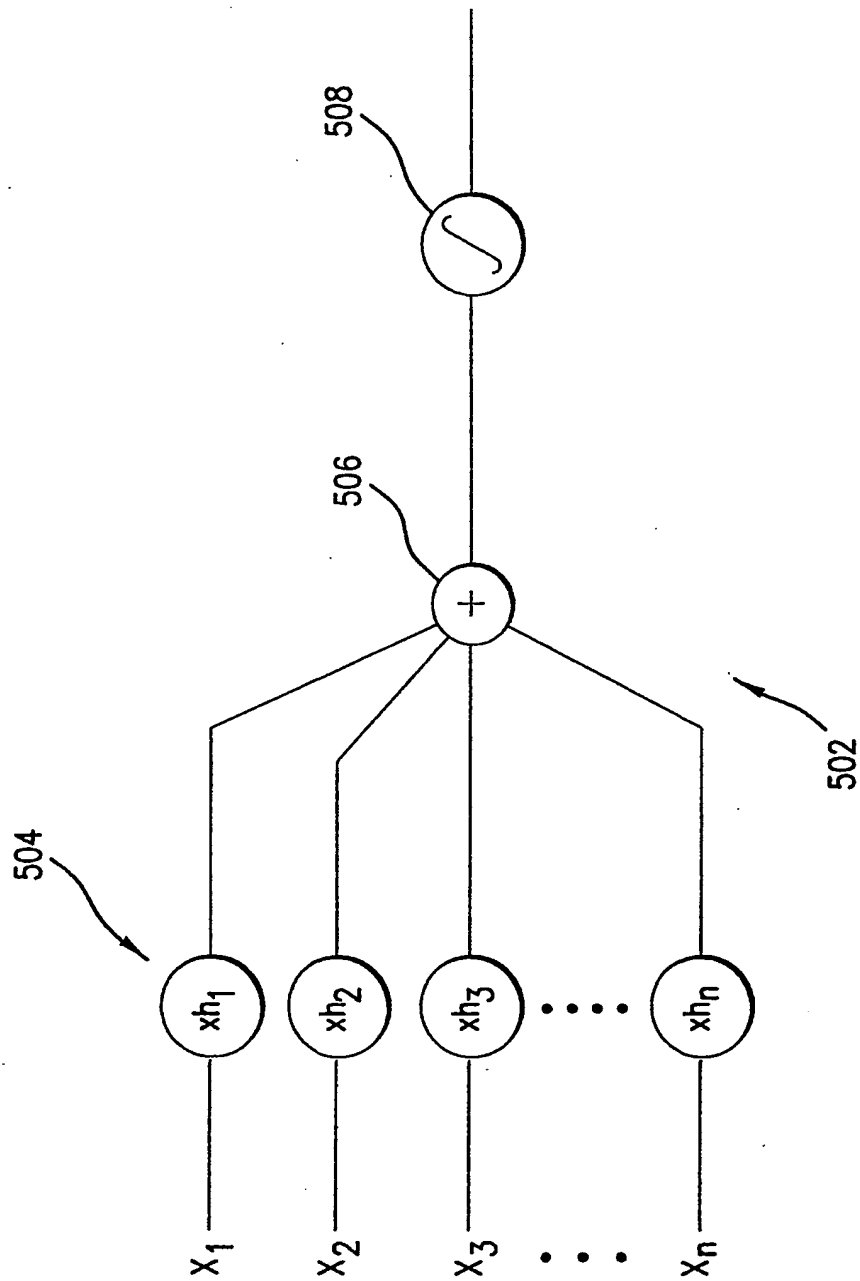


FIG.5

10/15

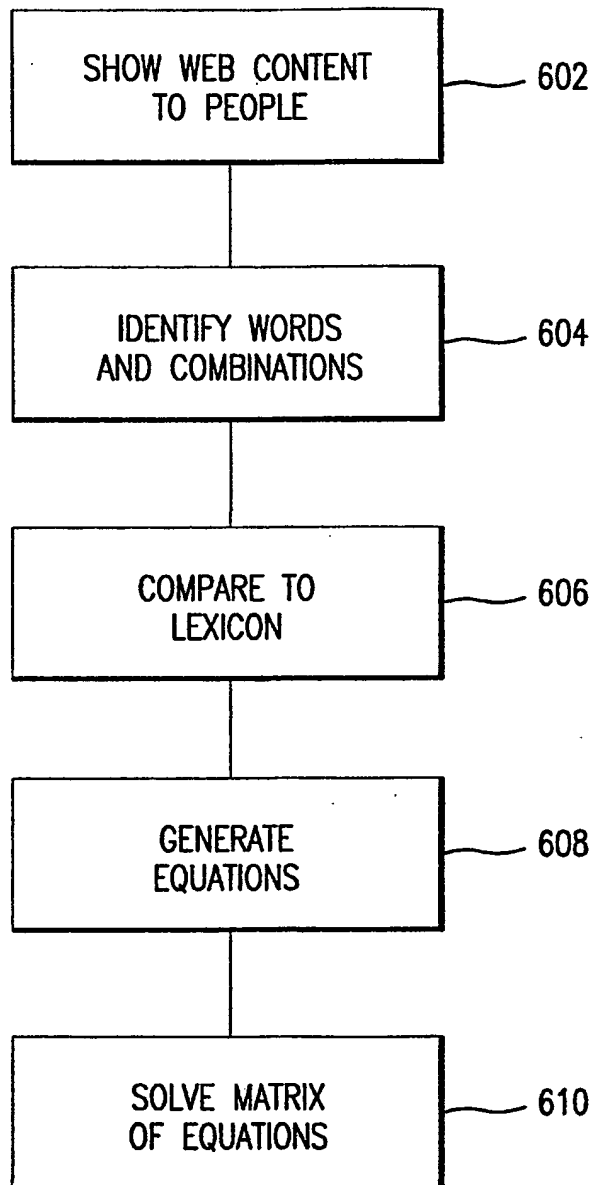


FIG. 6

SUBSTITUTE SHEET (RULE 26)

11/15

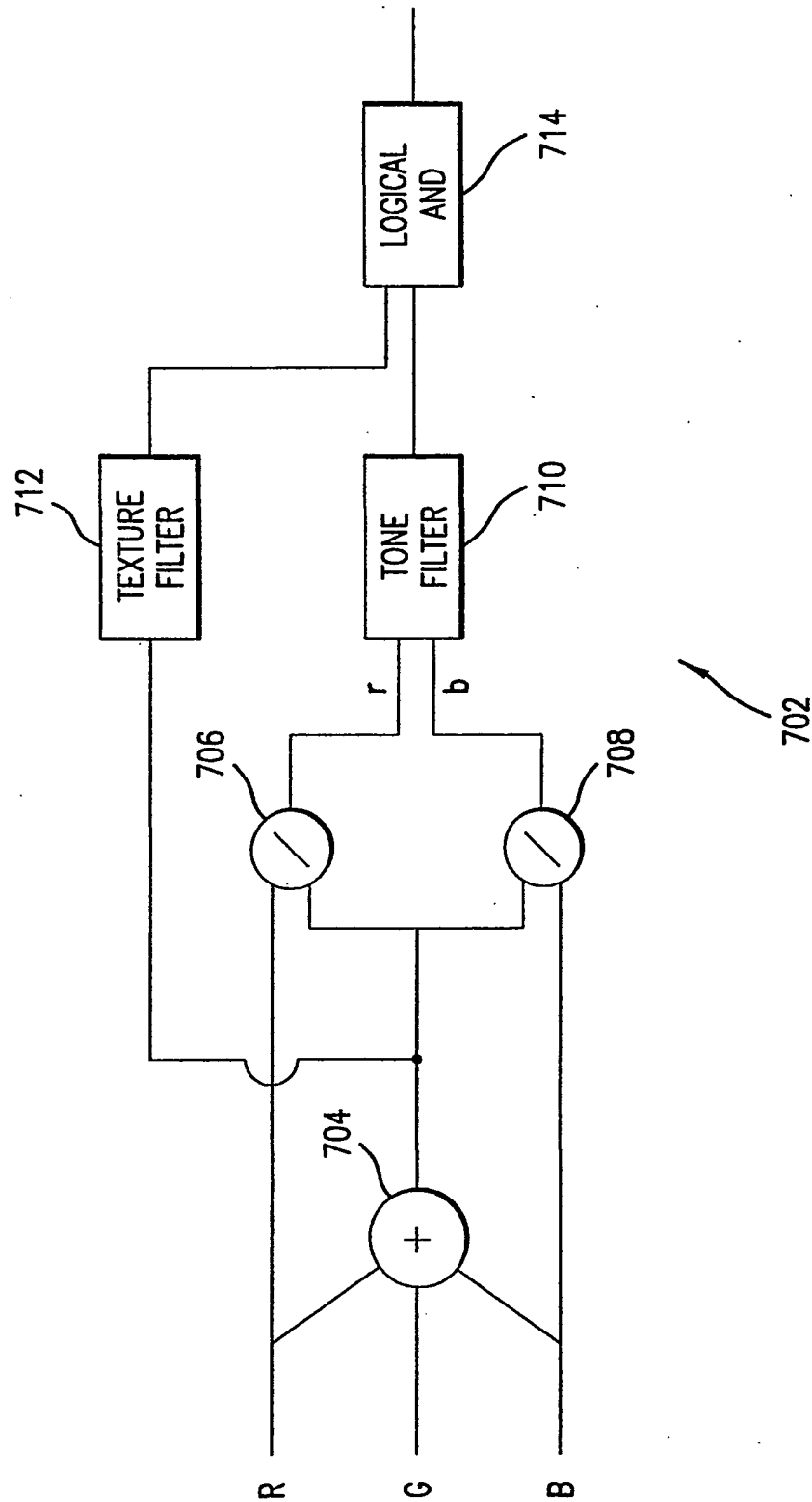


FIG.7

12/15

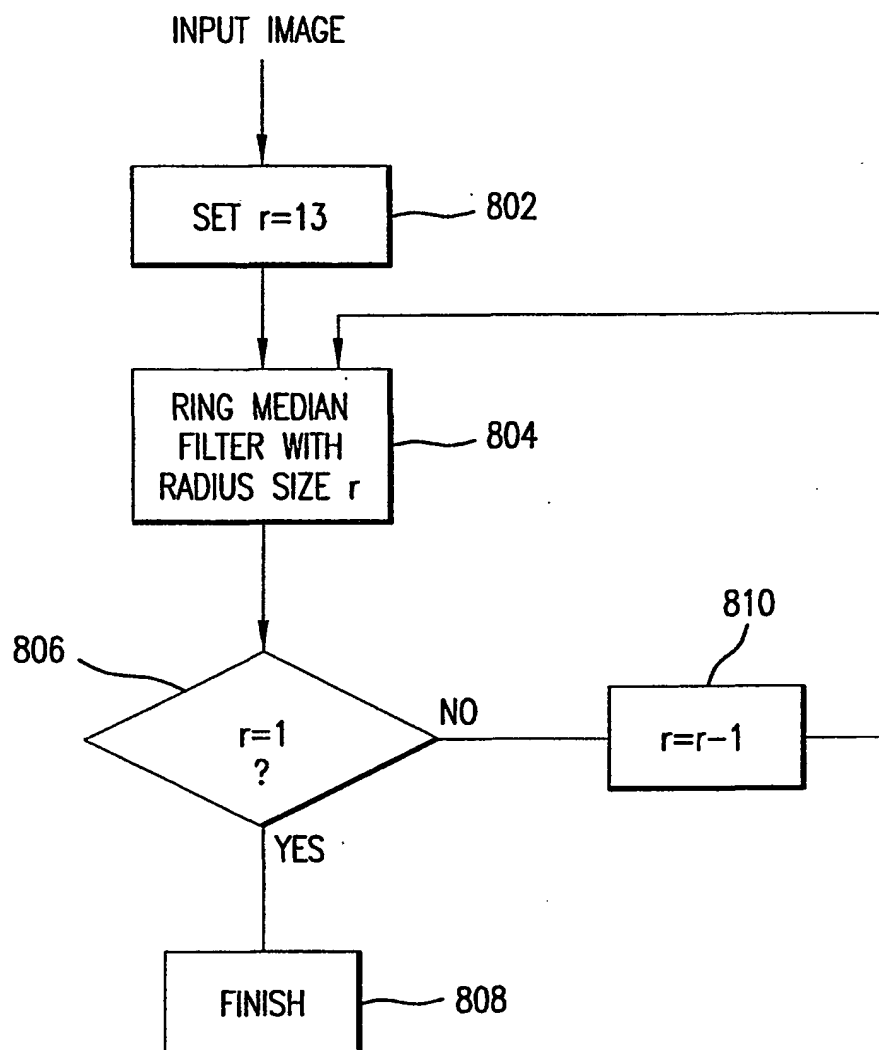


FIG.8A

13/15

			x_1			
	x_6				x_5	
x_2			x_k			x_0
	x_7				x_4	
			x_3			

FIG.8B

14/15

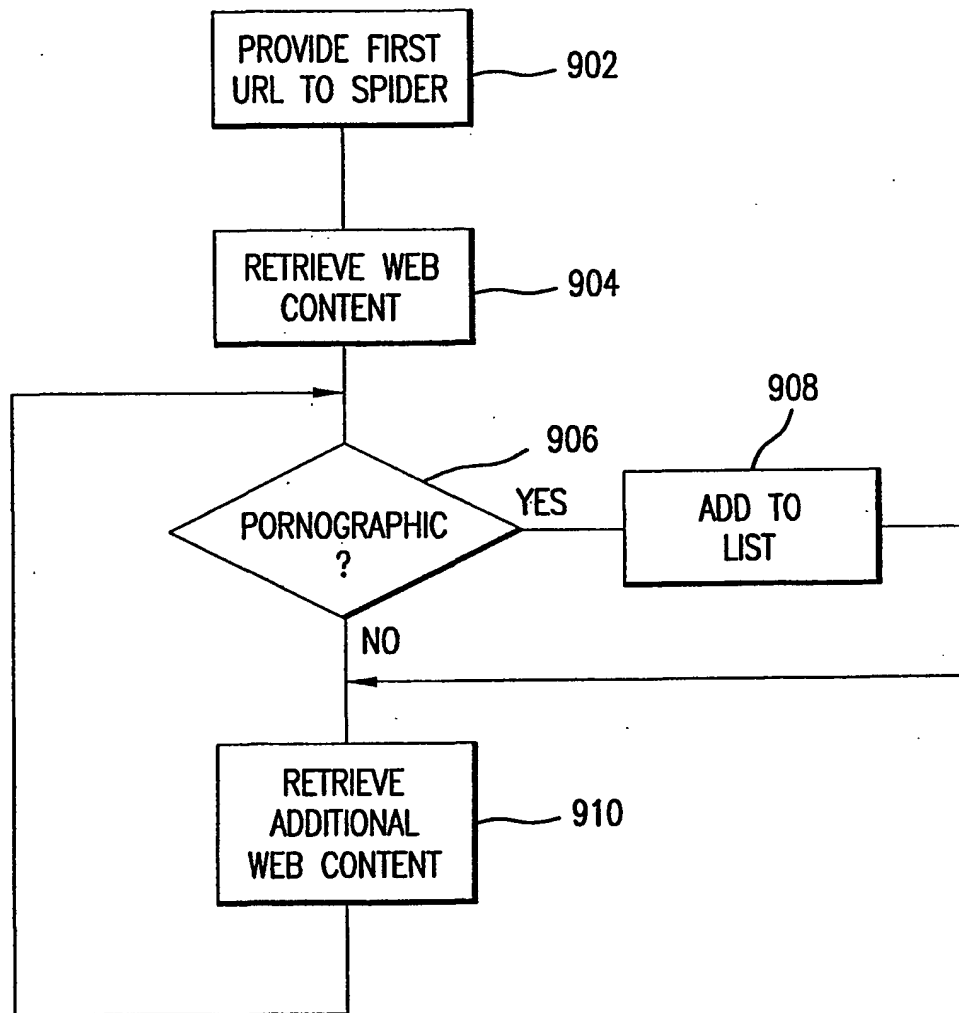


FIG.9

15/15

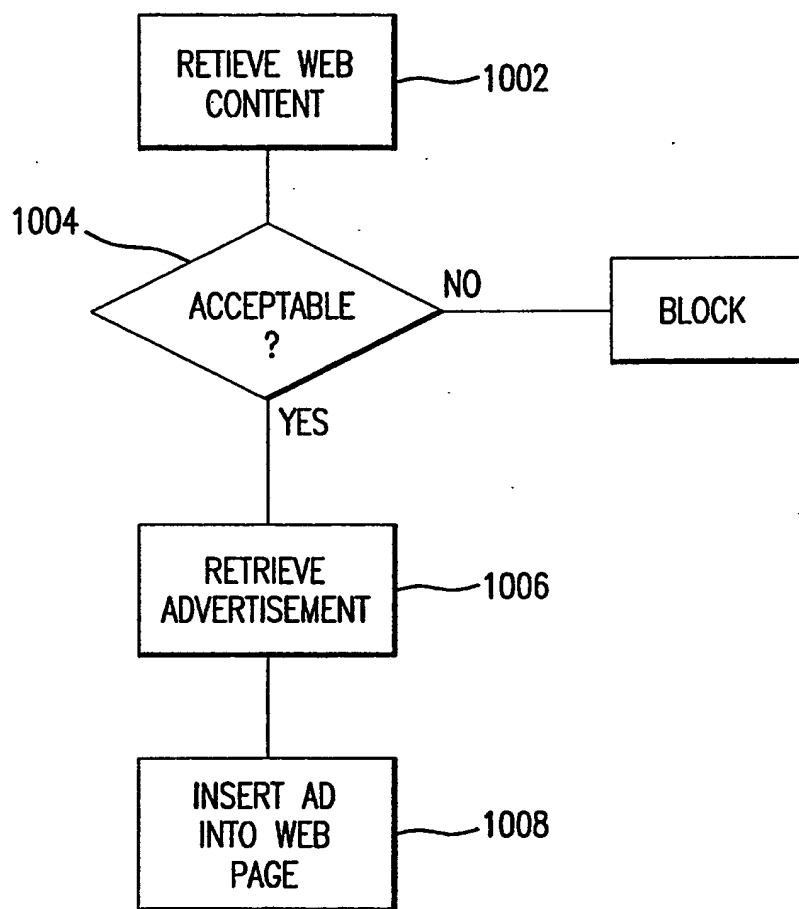


FIG.10

INTERNATIONAL SEARCH REPORT

International application No.
PCT/US01/05541

A. CLASSIFICATION OF SUBJECT MATTER		
IPC(7) : H04L 9/36 US CL : 713/200,201		
According to International Patent Classification (IPC) or to both national classification and IPC		
B. FIELDS SEARCHED		
Minimum documentation searched (classification system followed by classification symbols) U.S. : 713/200,201		
Documentation searched other than minimum documentation to the extent that such documents are included in the fields searched		
Electronic data base consulted during the international search (name of data base and, where practicable, search terms used) STN (USPAT, NLDB, INSPECT) fusion module, feature, text, image, filtering, web page, web site, pornographic		
C. DOCUMENTS CONSIDERED TO BE RELEVANT		
Category*	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No.
Y	US 5,839,103 A (MAMMONE et al) 17 November 1998, Fig. 3, col. 5, lines 39-64.	1-13
Y	"This.com Provides Free Web Filter Service to US Schools," Newsbytes, 28 October 1999, Newsbytes News Network, Newsletter.	1-13
Y	US 5,996,011 A (HUMES) 30 November 1999, Abstract, Col. 3, lines 23-40.	1-13
Y	US 5,941,944 A (MESSERLY) 24 August 1999, Fig. 8, Col. 6, lines 55-67.	6
Y	US 5,142,666 A (YOSHIZAWA et al) 25 August 1992, Abstract, Fig. 1., col. 1, lines 23-35.	10
<input checked="" type="checkbox"/> Further documents are listed in the continuation of Box C. <input type="checkbox"/> See patent family annex.		
* Special categories of cited documents:	"T"	later document published after the international filing date or priority date and not in conflict with the application but cited to understand the principle or theory underlying the invention
"A" document defining the general state of the art which is not considered to be of particular relevance	"X"	document of particular relevance; the claimed invention cannot be considered novel or cannot be considered to involve an inventive step when the document is taken alone
"E" earlier document published on or after the international filing date	"Y"	document of particular relevance; the claimed invention cannot be considered to involve an inventive step when the document is combined with one or more other such documents, such combination being obvious to a person skilled in the art
"L" document which may throw doubts on priority claim(s) or which is cited to establish the publication date of another citation or other special reason (as specified)	"&"	document member of the same patent family
"O" document referring to an oral disclosure, use, exhibition or other means		
"P" document published prior to the international filing date but later than the priority date claimed		
Date of the actual completion of the international search 01 JUNE 2001	Date of mailing of the international search report 26 JUN 2001	
Name and mailing address of the ISA/US Commissioner of Patents and Trademarks Box PCT Washington, D.C. 20231 Facsimile No. (703) 305-3230	Authorized officer LY V. HUA <i>James R. Matthews</i> Telephone No. (703) 305-9684	

INTERNATIONAL SEARCH REPORT

International application No.
PCT/US01/05541

C (Continuation). DOCUMENTS CONSIDERED TO BE RELEVANT

Category*	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No.
Y	"News Briefs: Sybase Adds Visual Searches," Document Imaging Report, (28 May 1997) Vol. 7, No. 11., Phillips Business Information, Inc.	13

INTERNATIONAL SEARCH REPORT

International application No.
PCT/US01/05541

Box I Observations where certain claims were found unsearchable (Continuation of item 1 of first sheet)

This international report has not been established in respect of certain claims under Article 17(2)(a) for the following reasons:

1. ☐ Claims Nos.:
because they relate to subject matter not required to be searched by this Authority, namely:

2. ☐ Claims Nos.:
because they relate to parts of the international application that do not comply with the prescribed requirements to such an extent that no meaningful international search can be carried out, specifically:

3. ☐ Claims Nos.:
because they are dependent claims and are not drafted in accordance with the second and third sentences of Rule 6.4(a).

Box II Observations where unity of invention is lacking (Continuation of item 2 of first sheet)

This International Searching Authority found multiple inventions in this international application, as follows:

Please See Extra Sheet.

1. ☐ As all required additional search fees were timely paid by the applicant, this international search report covers all searchable claims.
2. ☐ As all searchable claims could be searched without effort justifying an additional fee, this Authority did not invite payment of any additional fee.
3. ☐ As only some of the required additional search fees were timely paid by the applicant, this international search report covers only those claims for which fees were paid, specifically claims Nos.:
4. ☒ No required additional search fees were timely paid by the applicant. Consequently, this international search report is restricted to the invention first mentioned in the claims; it is covered by claims Nos.:

1-13

Remark on Protest

- ☐ The additional search fees were accompanied by the applicant's protest.
☐ No protest accompanied the payment of additional search fees.

INTERNATIONAL SEARCH REPORT

International application No.
PCT/US01/05541

BOX II. OBSERVATIONS WHERE UNITY OF INVENTION WAS LACKING

This ISA found multiple inventions as follows:

This application contains the following inventions or groups of inventions which are not so linked as to form a single inventive concept under PCT Rule 13.1. In order for all inventions to be searched, the appropriate additional search fees must be paid.

Group I, claim(s) 1-13, drawn to a system for identifying possibly pornographic web sites.

Group II, claim(s) 14-16, drawn to a method for inserting an advertisement into retrieved web content.

The inventions listed as Groups I and II do not relate to a single inventive concept under PCT Rule 13.1 because, under PCT Rule 13.2, they lack the same or corresponding special technical features for the following reasons: the apparatus as claimed in the claims of Group I is used to practice another and materially different process such as for identifying a pornographic web site, rather than a process for inserting an advertisement into retrieved web content.